Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1423 – 1430

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

Identifying Negative Interactions in Protein-Protein Interaction Network using Weak Edge-Edge Domination Set

Sminu Izudheen[a] , Sheena Mathew[b]

[a]*Assistant Professor, Department of Computer Science & Engineering, Rajagiri School of Engineering & Technology, Kerala, India, sminu_i@rajagiritech.ac.in*

[b]*Professor, Division of Computer Engineering, School of Engineering, Cochin University of Science & Technology, Kerala, India, sheenamathew@cusat.ac.in*

Abstract

Link prediction has recently attracted the attention of many researchers as an effective technique to understand the associations between proteins. But most of the work in this area was concentrated on predicting existence of links in future. Very few works has explored the prediction of links that might disappear in future. Also, links predicted by these methods may contain high levels of wrong interactions. In this paper, we propose a method to optimize the negative link predicted in protein network through Weak Edge-Edge Domination (WEED) set. We have tested our model using different standard prediction methods and the results obtained assert that our method can be used as an effective method to reduce false positive rate of negative links predicted in protein network.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

Keywords:Link Prediction; Protein Protein Interactions; Domination Set; Weak Edge Edge Domination Set.

1. Introduction

Protein interactions are important for numerous biological functions. For example, signal transduction, the process by which signals from exterior of a cell is mediated to interior of the cell is controlled by protein-protein interaction (PPI) of the signaling molecules. It also plays a fundamental role in many biological processes, including the pathway towards many diseases like cancer. Several efforts have been made to identify protein interactions, so that

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.167

biological systems can be understood better. The cost for experimentally detecting physically interaction between proteins in laboratory is very high and hence our current knowledge about protein networks is substantially incomplete [1,2]. Instead of blindly checking all possible interactions, perform prediction based on the observed interactions and then focusing on links most likely to vanish can sharply reduce the experimental costs [3]. This motivated us towards link prediction which is one of major computational problem in this area. Protein network is a complex network with proteins as nodes and their interactions as links. They are very dynamic objects, as they grow and change quickly over time through the addition of new edges. Protein network may always prompt some challenging questions like, how long a pair of proteins will remain connected or can a link disappear? What about the proteins that are not connected in the current state, is it possible that they will get connected sometime in the future? Link prediction problem in protein network has attracted much attention because, understanding the dynamics that drives the evolution of protein network is always challenging. However, researchers concentrated mostly on predicting how a protein network may grow by adding new links. Most of the previous works on link prediction is limited to the prediction of the links that will be added to the network during an interval of time. Predicting links that may be dropped from the network is still to be investigated, the paper discuss about this shrinking problem. Here we propose an efficient method to optimize the link predicted through Minimum Weak Edge-Edge Domination set[4] of the protein network, which will help to reduce the false positive rate in the negative links predicted. The results obtained assert that our method can be used as an effective method for link prediction in protein network.

2. Related Work

Protein-protein interactions (PPIs) are one of the most intensively analyzed networks in biology and there are a multitude of biochemical and biophysical methods to detect them [5,6]. Since molecular biology techniques used are very expensive and time-consuming, researchers depend on graph theory techniques to study them.

Nantia Iakovidou et.al.[7] uses a multiway spectral clustering analysis, a technique that uses information obtained from the top few eigenvectors and eigenvalues of the normalized laplacian matrix as a method to predict links in PPI network. W. Pentney et.al. [8] prove that their algorithm applying spectral clustering offers competitive performance on sequence data. A simple and unified derivation of spectral clustering of biological data is presented in [9]. A tool for the identification of PPIs, which can be used to detect interactions across the entire proteome of an organism is given in [10]. Local Protein Community Finder is a tool developed by authors on [11] to find community close to a queried protein in any network specified by the user. To predict protein interactions in yeast network Y. Yamanishi et.al. [12] introduced a method based on variant of kernel canonical correlation analysis.

Link prediction has also attracted researchers from the area of social networking. Commonly, two nodes are more likely to be connected if they are more similar. A Comparison between similarity indices is presented in [13], where node-dependent indices like Common Neighbors[14], Jaccard coefficient [15], Adamic-Adar Index [16], Preferential Attachment [17] and path-dependent indices like Katz Index [18],Hitting Time [19], Commute Time [20], Rooted PageRank [21], SimRank [22] and Blondel Index [23] were considered. Zhou et al.[24] proposed Resource Allocation index and Local Path index as a measure to compare two nodes. Results shows that the local path index provides much accurate prediction compared with the global index[25]. On a weighted network, weak links play an important role than strong links[26]. Likelihood for the existence of a link between two nodes was estimated through local path index in [27]. Weiping Liu et.al. [28] present a method to find node similarity based on local random walk. They illustrate that the method has lower computational complexity compared with other random-walk-based similarity indices, such as average commute time (ACT) and random walk with restart (RWR).

Researchers were mainly concentrated on predicting how a protein network may grow by adding new links. Very few works have addressed the problem of predicting links that may be dropped from the network in future. Wadhah Almansoori et. al. [29] present a method to find the negative links from the positive links predicted. They have applied the model to two different domains, namely health care and stock market. Yuan Zhu et. al.[30] presents a generative network model to identify both spurious and missing interactions in a protein network. In this paper we propose a method to optimize the result predicted using any similarity index through all possible minimum WEED-set of the network. We represented protein interactions as an undirected graph and predicted the links that may disappear in future using various standard methods viz, Common Neighbors (CN), Jaccard coefficient (JC), Adamic-Adar Index (AA), Preferential Attachment (PA), Local Random Walk (LRW) and Superposed Random Walk(SRW). We then

optimize the links predicted by calculating the Weak Edge-Edge Domination (WEED) set of the predicted links. When compared the result with standard methods, optimization performed using WEED set shows significant improvement. This asserts that link prediction can be improved through minimum WEED set of the network.

3. Methods and Data

3.1. Data

For the present study protein-protein interaction data is downloaded from MINT [31] database. After removing redundancy and self interactions we had 187455 protein interactions among 12119 proteins. To know more about the data, degree distribution was plotted. Figure 1 shows that dataset follows a skewed distribution with degree ranging from 1 to 600 and the skewness value is 4.64. Most nodes have relatively small degree, only few are with very large degree which forms long tail in the distribution. These large degree nodes form possible hubs in the network. The scale free property of the PPI network is evident from the figure. Scale free property means that the degree distribution approximate to power law. i.e., the probability that a node has k links follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent.

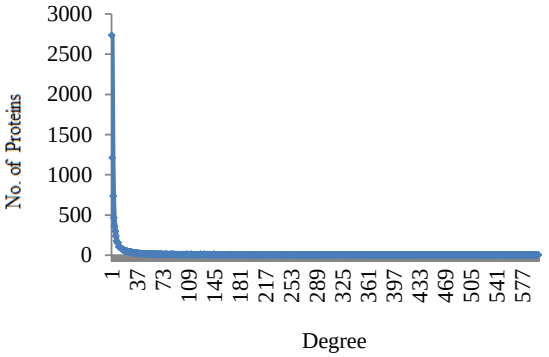

Fig. 1. Degree distribution of MINT dataset

Since the number of interactions is very huge, sampling is done by randomly selecting interactions, ensuring that the degree distribution is not disturbed. To test the performance of the algorithm a test data set was generated from the sampled data with an assumption that the network follows a Gaussian distribution. A connection is added or removed from the network based on a Gaussian probability value. For this, k random proteins p_{ref} are selected from the sampled data set and Mahalanobis distance, d towards all proteins p_{cur} within a given circumference from p_{ref} was calculated. The probability value of the protein p_{cur} with respect to the reference protein p_{ref} will then be $p = e^{-d}$. If the probability p is greater than a random function the connection between p_{ref} and p_{cur} is toggled. To ensure that the data set generated follows the same pattern as the sampled dataset, its degree distribution was plotted and compared with the sampled data set. The degree distribution of the sampled dataset and the test data set is given in Fig. 2(a). and 2(b). respectively.

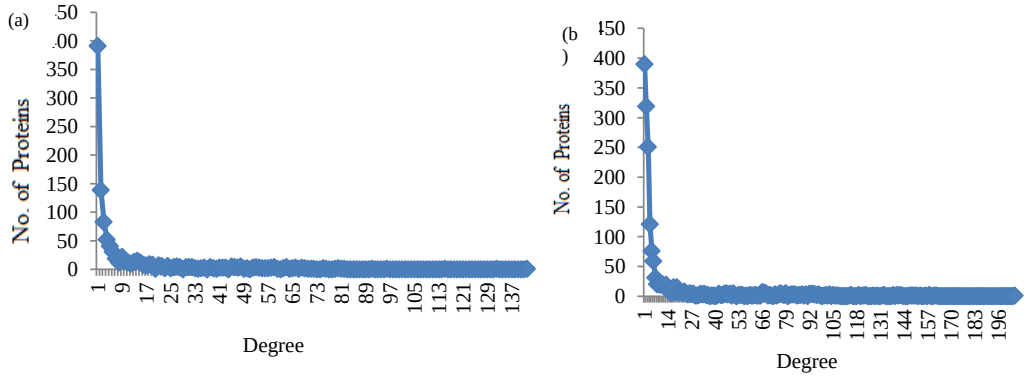

Fig. 2. (a) Degree distribution of sample dataset; (b) Degree distribution of test dataset

3.2. Link Prediction based on Similarity Index

Consider an undirected graph G(V,E), where V is the set of vertices and E represents the set of edges. Two nodes are more likely to be connected if they are similar. A brief introduction about various similarity indices are given below.

3.2.1. Common Neighbour (CN)

Two nodes, x and y, are more likely to have a link if they have more common neighbors. One measure by which we can express this neighborhood overlap is

$$S_{xy}CN = |k_x \cap k_y| \tag{1}$$

where k_x ,k_y represents neighbors of *x* and *y* respectively.

3.2.2. Jaccard Coefficient (JC)

Jaccard Coefficient is defined as the size of the intersection divided by the size of the union of the sample sets

$$S_{xy}JC = (|k_x \cap k_y|)/(|k_x \cup k_y|) \tag{2}$$

3.2.3. Adamic Adar (AA)

This method computes the similarity between any two vertices x and y using a common feature of the two. The similarity measure is then

$$S_{xy}AA = \sum_{Z \in (kx \cap ky)} \frac{1}{\log(Z)} \tag{3}$$

3.2.4. Preferential Attachment (PA)

Preferential Attachment is defined by the product of two related nodes' degrees or summarization of their degrees. i.e., the pairwise interaction between nodes x and y is proportional to k_x ,k_y which represents neighbors of x and y respectively.

$$S_{xy}PA = |k_x| . |k_y| \tag{4}$$

3.2.5. Random Walk (RW)

Probability that a random walker starting at node x will move to y in the next step is represented by transition probability matrix, P, with $P_{xy} = a_{xy}/k_x$, where a_{xy} equals 1 if node x and node y are connected, 0 otherwise, and k_x

denotes the degree of node x. The probability that a random walker located at node x will be located at node y after t steps is given by

$$\pi_x(t) = P'.\pi_x(t\text{-}1) \tag{5}$$

where $\pi_x(0)$ is an Nx1 matrix with x = 1 and all other values are 0's and P' is the transpose matrix. The similarity between node x and node y on Local Random Walk (LRW) [27] is given by

$$S_{xy}\,LRW(t) = \frac{k_X}{2|E|}.\pi_{xy}(t) + \frac{k_X}{2|E|}.\pi_{yx}(t) \tag{6}$$

Since we are considering an undirected graph, 2|E| represents the number of links in the network.

As the random walk based similarity measure is that it shows sensitive dependence to sub graph away from nodes x and y, even when x and y are connected by short paths [32]. Hence, the probability for the random walker to go farther from x and y, even though they are close to each other is high. But proteins have a tendency to connect with ones nearby rather than far way. This may lead to low prediction accuracy. To solve this problem we can continuously release the walkers at the starting point. Hence there will be higher similarity between target node and nearby nodes. By superposing the contribution of each walker, we get the next similarity index, Superposed Random Walk (SRW).

$$S_{xy}\,SRW(t) = \sum_{l=1}^{t} S_{xy}\,LRW(l) \tag{7}$$

3.3. Minimum Weak Edge-Edge Domination Set

Let *u* and *v* be any two adjacent vertices of an undirected graph G(V,E). The open neighborhood of a vertex, u∈ V is N(u) = { v∈V| uv∈E} and the closed neighborhood is N[u] = N(u) ∪ {u} [33]. The vertex *u* weakly dominates *v* if deg (*u*) ≤ deg(*v*) [34] and a set D⊆V is a *weakly dominating set* [WD-set], if every v∈V-D is weakly dominated by some u∈D.

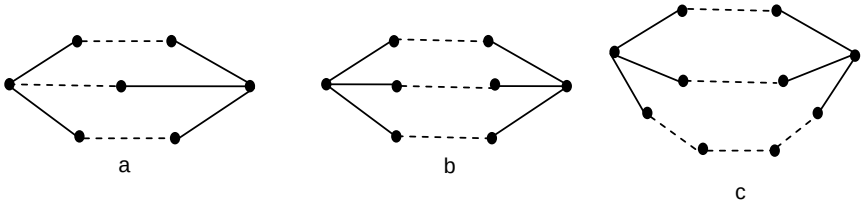

Fig. 3. Example showing minimum WEED set of various undirected graphs. Edges in minimum WEED set is represented as dotted lines

The concept of domination set on vertices has been extended to edges in [4]. In [4] the degree of an edge x=uv is defined as the number of edges adjacent to the edge *x*, given by deg(*x*) = deg (*u*) + deg (*v*) – 2. Also, the open neighborhood of an edge x=uv is N(x) = {y∈E | y is adjacent to x} and the closed neighborhood is N[x] = N(x) ∪ {x}. An edge x weakly *e*-dominates an edge y if y∈N[x] and deg(x) ≤ deg(y). A set D⊆E is a Weak Edge-Edge Dominating set [WEED-set], if every edge in E-D is weakly *e*-dominated by an edge in D. A set L is a minimal WEED set of G if, and only if, for any x∈L, either of the following two conditions holds.(i) No edge in L weakly e-dominates the edge x. (ii) There exists an edge y∈E-L which is uniquely weakly e-dominated by the edge x. Minimum WEED set for various undirected graph is shown in Fig. (3).

4. Algorithm

Protein interactions are represented as an undirected graph G(V,E), where V represents the set of proteins and E the set of interactions between them. In this protein network, negative links are predicted based on the assumption

that the interaction between two proteins is more likely to get dropped in future if they are less similar. The system is trained to generate similarity score for every pair of nodes. The similarity can be calculated using any of the similarity measures viz, Common Neighbors (CN), Jaccard coefficient (JC), Adamic-Adar Index (AA), Preferential Attachment (PA), Local Random Walk (LRW) and Superposed Random Walk(SRW). Now based on similarity score, sort the existing links in ascending order. The links which are in the top of the list are more likely to get dropped.

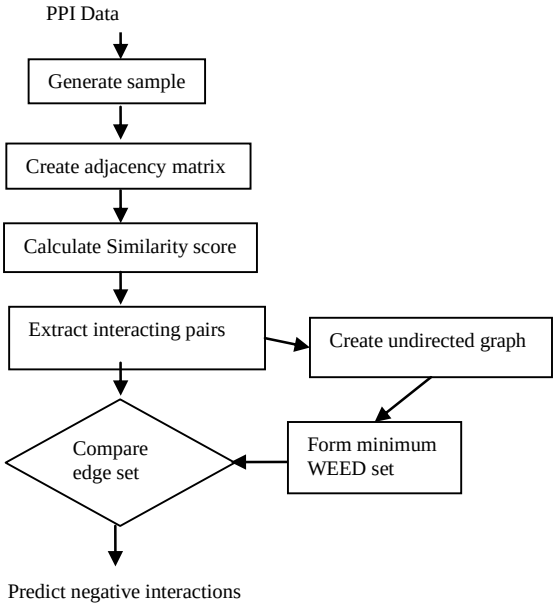

Fig. 4. Schematic overview of Negative Link Prediction in Protein Networks

```
Select links predicted using similarity score
Create the adjacency matrix for the links predicted, A
Find degree of each edge as From_degree + To_degree – 2
Sort non zero degree edges in asccending order of degree, E
Select edges with lowest degree to E'
while()     // exit after finding WEED starting with all lowest degree
            //edges
{
     Unmark all edges in E
     Select the top edge from E' to T
     while()   //exit either WEED is found or no WEED possible
     {
       Find neighbors of T and mark it
       Mark current node
       Add T to WEED set
       If all edges are marked, WEED found, break
       Set T as neighbor with minimum degree
        If neighbor also not present, then WEED not possible, break
      }
    If WEED not possible with E', add edges with next min. degree to E'
     If WEED is found, remove edges in WEED set from E'
      Found WEED and E' is empty, break
}
```

Fig. 5. Algorithm to find all possible WEED set

A minimum WEED set of a graph represents the set of edges which weakly dominate the rest of edges in the graph. Hence, the edges in minimum WEED set are more likely to get dropped in future. Using this concept, the above predicted result can be optimized by finding the minimum WEED set. For this, the predicted links are represented as an undirected graph and all possible minimum WEED-set of the graph is generated. Since edges in the minimum WEED set represent weak connection in the network, these edges are more likely to get dropped. Schematic overview of the method is given in Fig. 4.

The most challenging part here is to calculate minimum WEED set of a graph. An algorithm to calculate all possible WEED set is given in Fig. 5. An undirected graph is created by extracting links predicted and all the non zero degree edges are sorted in the ascending order of degree. Edges are then processed in this order. The algorithm will exit from the outer while loop after finding all possible WEED set starting with highest degree edges. The inner while loop will check whether a WEED set is possible or not. If it is possible, it will return the set. Every step inside the inner while can be computed in not more that O(n) time, where n represents the number of proteins. Hence the asymptotic complexity of the algorithm will be $O(n^3)$.

5. Result and Discussion

From the protein-protein interaction data a protein network is created and represented as an adjacency matrix. Links which are probable to get dropped in future is predicted using different similarity measures viz, Common Neighbors, Jaccard coefficient, Adamic-Adar Index, Preferential Attachment, Local Random Walk and Superposed Random Walk. In this paper we propose a method to optimize the result predicted using any similarity index through all possible minimum WEED-set of the network. From the results obtained the following observations are noted. To quantify the accuracy of the prediction algorithm, two standard metrices, AUC and precision were calculated. Table 1 gives the AUC and precision for various similarity indices. It may be noted that for all similarity measure, the prediction can be improved through the calculation of WEED set. In the proposed method, improvement in accuracy during prediction is achieved due to the reduction in false positive rate.

Table 1. AUC and Precision for various Similarity Indices

Similarity Index	AUC	Precision
CN	0.615	0.32
CN with WEED	0.63	0.33
JC	0.6	0.2
JC with WEED	0.6	0.24
AA	0.615	0.32
AA with WEED	0.632	0.33
PA	0.65	0.32
PA with WEED	0.68	0.33
LRW	0.7	0.72
LRW with WEED	0.725	0.78
SRW	0.675	0.64
SRW with WEED	0.69	0.67

6. Conclusion

This paper presents a method for predicting negative interactions from a PPI network. The WEED algorithm presented in the paper can effectively reduce the false positive rate in predicting negative interactions using any similarity index. The experiments are implemented on a simulated data set extracted from MINT dataset, assuming the data set follows a Gaussian distribution. The result obtained indicates that it is effective to evaluate weak interactions on a PPI network.

References

[1] N. D. Martinez, B. A. Hawkins, H. A. Dawah, B. P. Feifarek. Effects of sampling effort on characterization of food-web structure. Ecology 1999; 80:1044-55.

[2] E. Sprinzak, S. Sattath, H. Margalit,. How reliable are experimental protein—protein interaction data?.Journal of Molecular Biology 2003; 327(5): 919-23.

[3] A. Clauset, C. Moore, M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. Nature 2008; 453: 98-101.

[4] R. S. Bhat, S. S. Kamath, Surekha R. Bhat. Strong (Weak) Edge-Edge Domination Number of a Graph. Applied Mathematical Sciences 2012; 6:5525 – 5531.

[5] T. Kocher and G. Superti-Furga. Mass spectrometry based functional proteomics: from molecular machines to protein networks. Nature Methods 2007; 4: 807-15.

[6] L. Liua, Y. Caic, W. Lua, K. Fenge, C. Penga and B. Niu. Pre-diction of protein-protein interactions based on PseAA composition and hybrid feature selection. Biochemical and Biophysical Research Communications 2009; 380: 318-22.

[7] Nantia Iakovidou, Panagiotis Symeonidis and Yannis Manolopoulos. Multiway Spectral Clustering Link Prediction in Protein-Protein Interaction Networks. IEEE EMBS International Conference on Information Technology Applications in Biomedicine 2010; 1-4

[8] W. Pentney and M. Meila. Spectral clustering of biological sequence data. 12th National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania 2005; 845-50.

[9] D. J. Higham, G. Kalna and M. Kibble. Spectral clustering and its use in bioinformatics. Journal of Computational and Applied Mathematics 2007; 204: 25-37.

[10] U. Stelzl, U. Worm, M. Lalowski, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005; 122: 957-68.

[11] K. Voevodski, S. Teng and Y. Xia. Finding local communities in protein networks. BMC Bioinformatics 2009; 10: 297-310.

[12] Y. Yamanishi, J.P. Vert and M.Kanehisa. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 2004; 20: 363-70.

[13] D. Liben-Nowell, J. Kleinberg. The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. &.Technol. 2007; 58(7): 1019-31.

[14] F. Lorrain, H. C. White. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1971; 1: 49-80.

[15] P. Jaccard. Etude de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Societe Vaudoise des Science Naturelles 1901; 37(142): 547-79.

[16] L. A. Adamic, E. Adar. Friends and neighbors on the Web. Social Networks 2003; 25: 211-30.

[17] A.-L. Barab´asi, R. Albert. Emergence of Scaling in Random Networks. Science 1999; 286: 509-12.

[18] L. Katz. A new status index derived from sociometric analysis. Psychmetrika 1953; 18: 39-43.

[19] F. Gobel, A. Jagers. Random walks on graphs. Stochastic Processes and Their Applications 1974; 2(4) : 311-36.

[20] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens. Random-Walk Computation of Similarities between Nodes of a Graph, with Application to Collaborative Recommendation. IEEE Trans. Knowledge and Data Eng. 2007*;* 19(3):355-69.

[21] S. Brin, L. Page. The Anatomy of a Search Engine. Comput. Netw. ISDN Syst. 1998; 30: 107-117.

[22] G. Jeh, J. Widom. SimRank: A Measure of Structural-Context Similarity. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York 2002; 271-79.

[23] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P.V. Dooren. A measure of similarity between graph vertices: Application to synonym extraction and Web searching. SIAM Rev 2004; 46(4): 647-66.

[24] T. Zhou, L. L¨u, Y.-C. Zhang. Predicting missing links via local information. The European Physical Journal 2009; 71: 623-30.

[25] L. L¨u, C.-H. Jin, T. Zhou. Similarity index based on local paths for link prediction of complex networks 2009; Phys. Rev. E 80;046122.

[26] L. L¨u, T. Zhou. Link prediction in weighted networks: The role of weak ties. Europhysics Letters 2010; 89: 18001.

[27] L. Lu, C. Jin and T. Zhou. Similarity index based on local paths for link prediction of complex Networks. Physical Review E 2009; 80(4): 046122.

[28] Weiping Liu and Linyuan LU, Link Prediction Using Local Random Walk. Europhysics Letters 2010; 89 (5): 58007.

[29] Wadhah Almansoor, Shang Gao, Tamer M. Jarada, Reda Alhaj and Jon Rokne. Link Prediction and Classification in Social Networks and its Application in Healthcare. IEEE IRI 2011; August 3-5, 2011. Network Modeling Analysis in Health Informatics and Bioinformatics 2012; 1: 27-36.

[30] Yuan Zhu, Xiao-Fei Zhang, Dao-Qing Dai, and Meng-Yun Wu. Identifying Spurious Interactions and Predicting Missing Interactions in the Protein-Protein Interaction Networks via a Generative Network Model. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2013; 10: 219-25.

[31] MINT: the Molecular INTeraction database, http://mint.bio.uniroma2.it/mint/Welcome.do.

[32] D. Liben-Nowell, J. Kleinberg. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 2007; 58(7): 1019-31.

[33] Razika Boutrig and Mustapha Chellali. A note on relation between the Weak and Strong Domination Numbers of a graph. Opuscula Mathematica 2012; 32: 235-38.

[34] E.Sampathkumar and S.S.Kamath. Mixed Domination in Graphs. Sankhya 1992; 54: 399-402.